

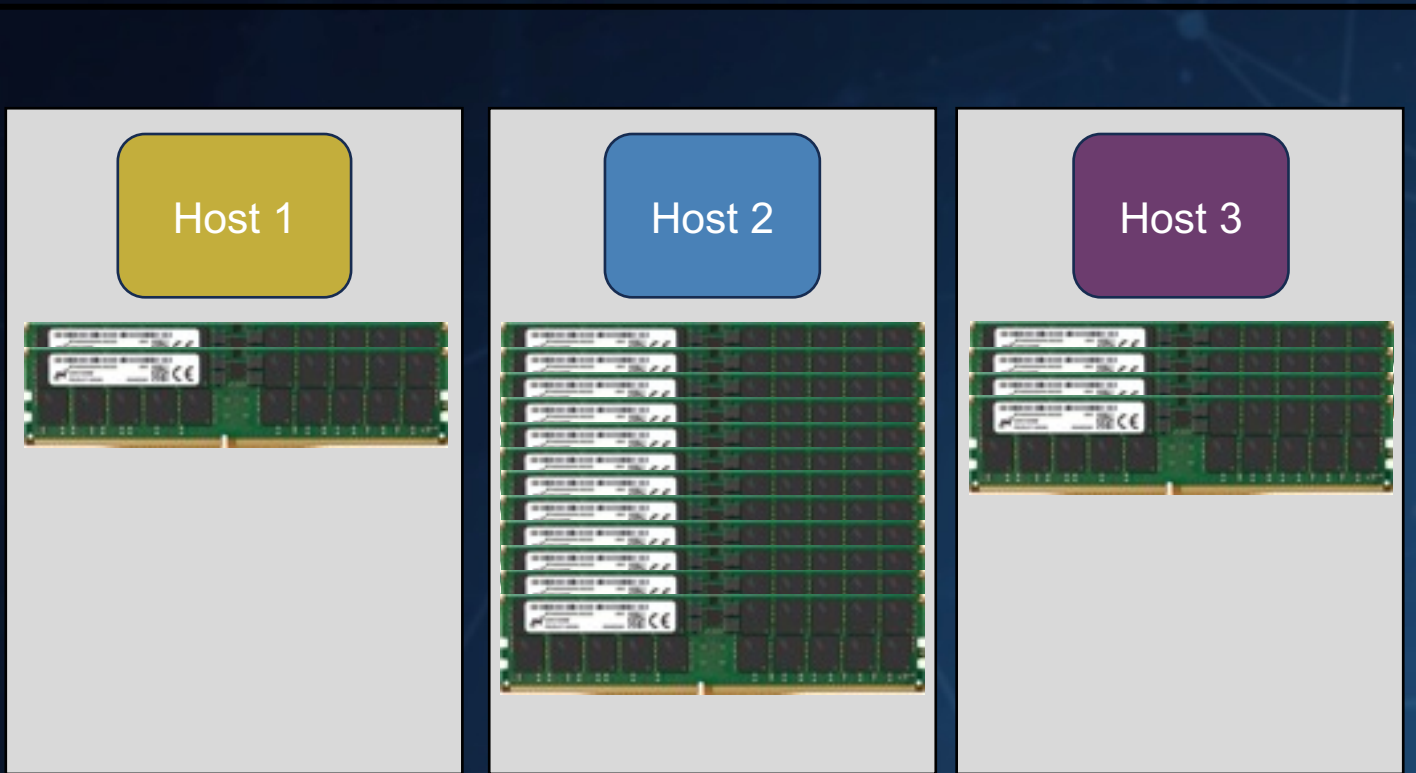
CXL for Composable and Disaggregated AI Computing

- XConn Technologies

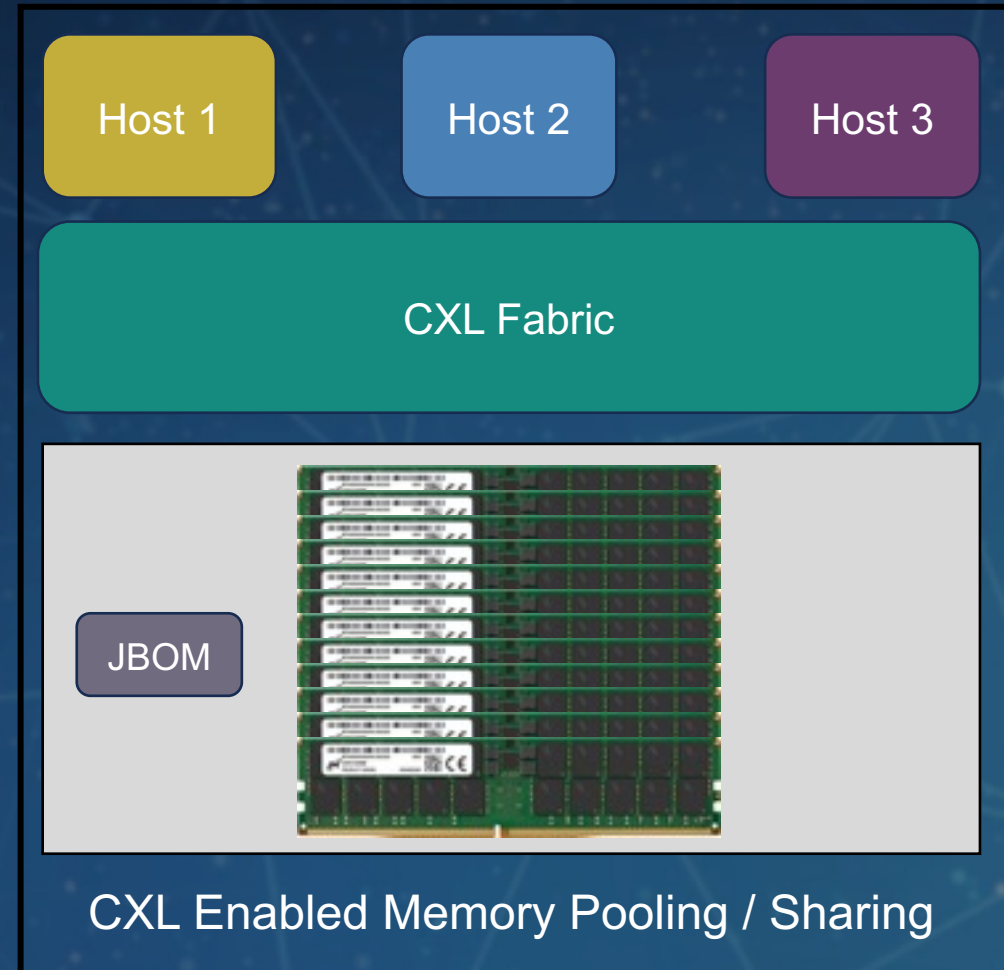
November 2023



CXL Enables DRAM Disaggregation for Usage Optimization



Multiple Heterogenous Servers
Each With Dedicated DRAM



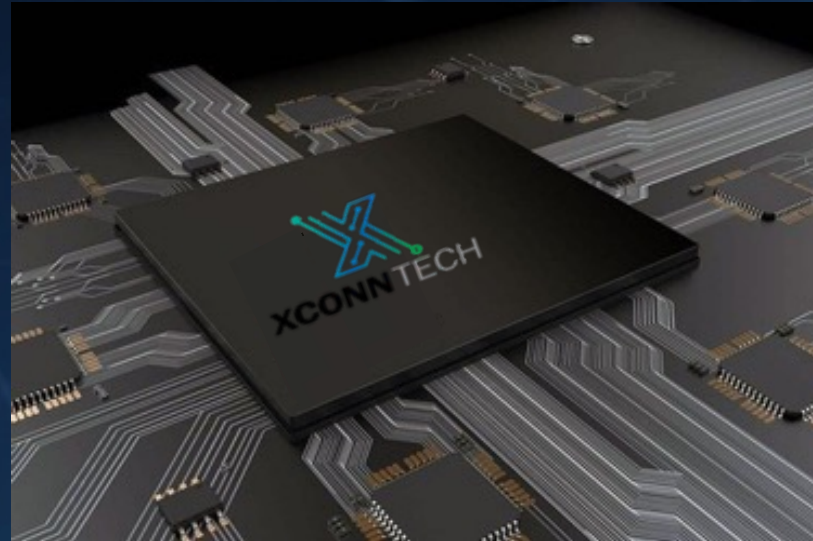
CXL Enabled Memory Pooling / Sharing

CXL Switch for Scalable Disaggregation



XConn Tech has the world's first CXL2.0 (XC50256) & PCIe 5.0 (XC51256) switch IC

2,048 GB/s total BW with 256 lanes



Lowest port-to-port latency

Lowest power consumption/port

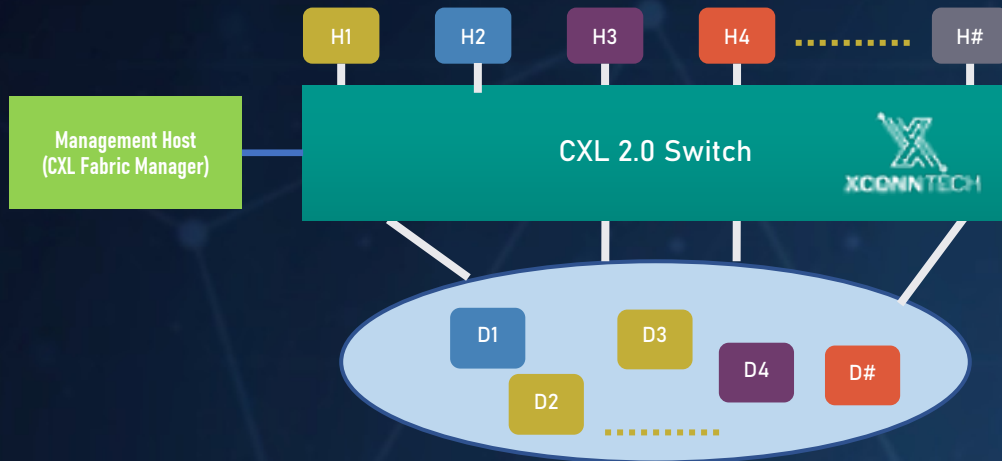
Reduced PCB area
Lower TCO

- Works with CXL 1.1 server processors, CXL memory devices
- Future compatible with the upcoming CXL 2.0 processors
- Works in hybrid mode (CXL/PCIe mixed)
- CS (customer samples) available now, MP 2Q24

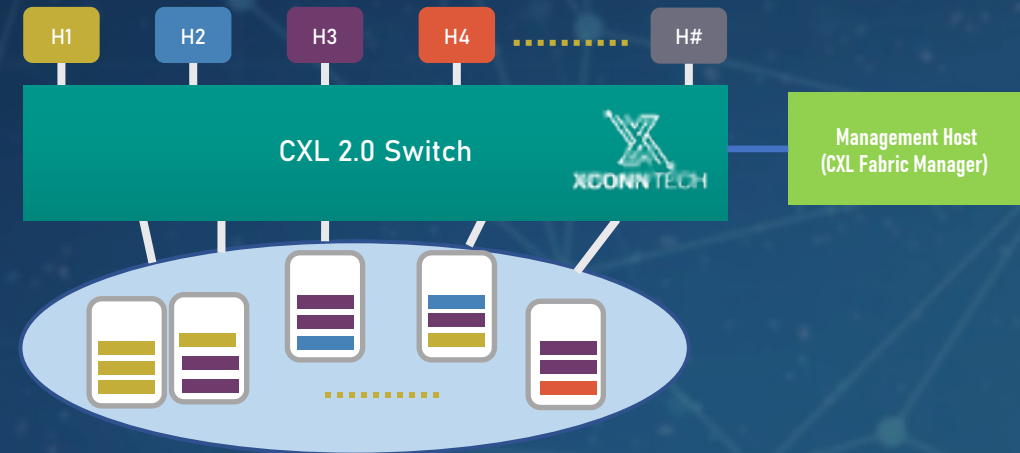
Scalable Memory Pooling & Sharing Enabled by CXL 2.0 Switch



Memory Pooling/Sharing with CXL 1.1 Hosts and Single Logical Devices

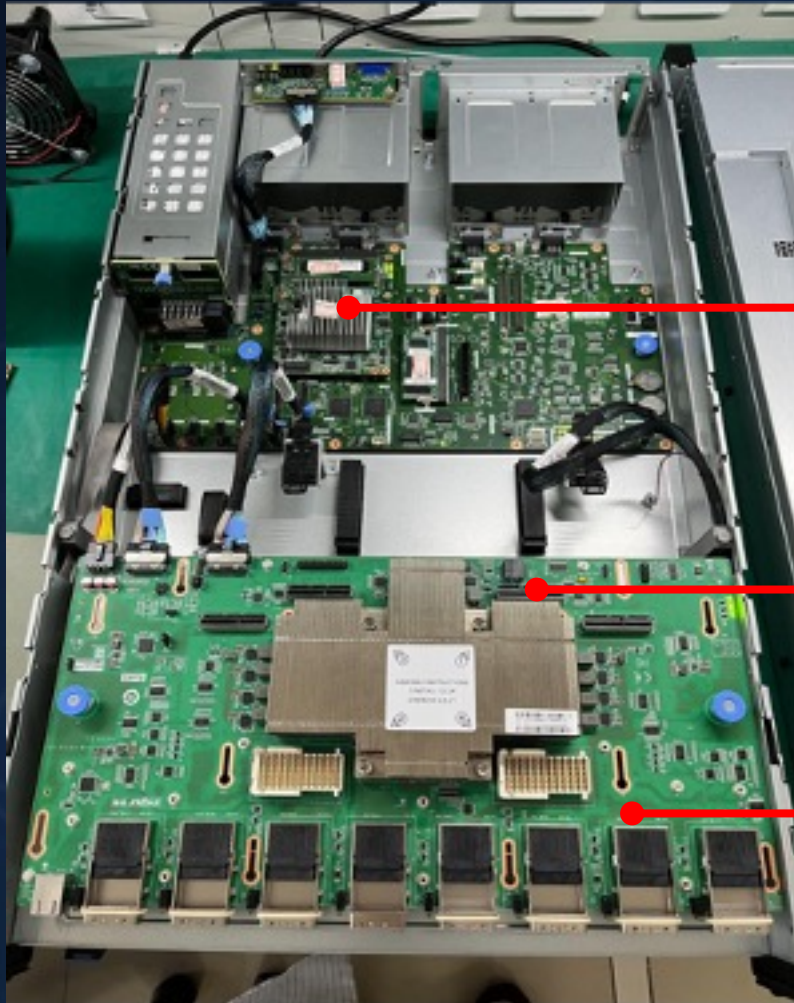


Memory Pooling with CXL 2.0 Hosts and Multiple Logical Devices



- One single XC50256 connects to 32 combined hosts/devices
- Fully support CXL Fabric Manager
- Support switch cascading for a larger size memory pool

Composable Disaggregated Memory System Enabled by 256-lane CXL Switch



x86 mCPU

Configure the CXL switch and memory module

CXL switch (256 lanes)

Connect the hosts and CXL memory modules

CDFP connector

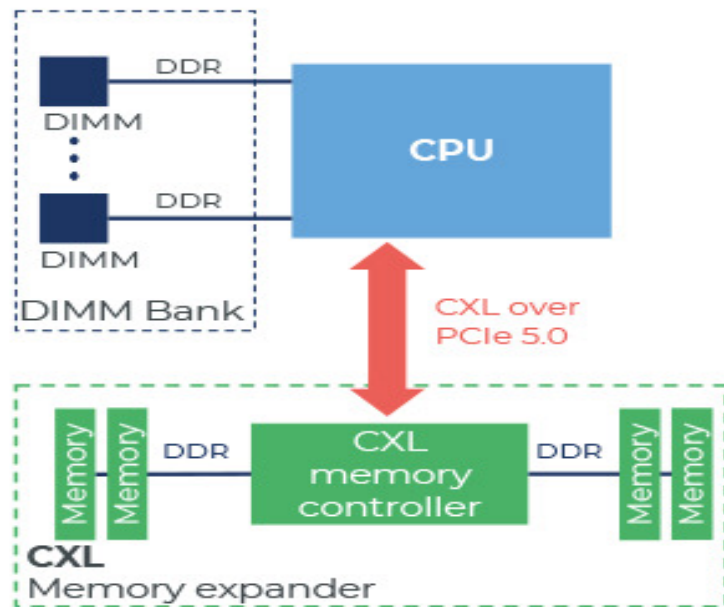
Connect to the hosts and CXL memory modules by CDFP cables

The Evolution of CXL Technology

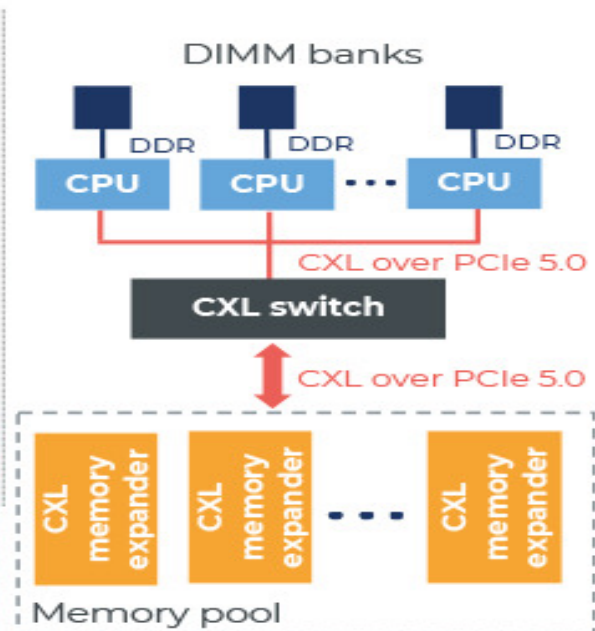
Memory expansion, pooling, and disaggregation using CXL integration

(Source : Memory-Processor Interface 2023 - Focus on CXL, Yole Intelligence, September 2023)

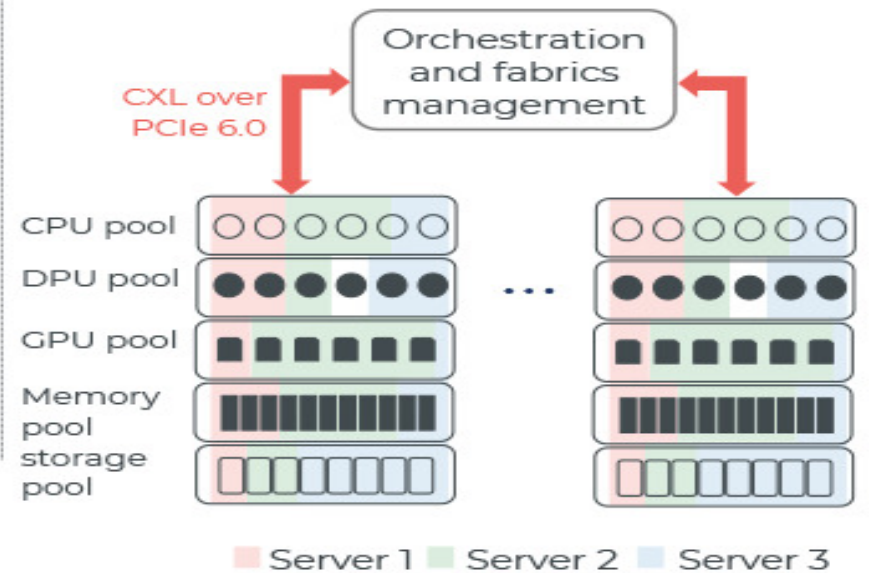
CXL 1.1 In-server memory expansion (server level)



CXL 2.0 Memory pooling (rack level)

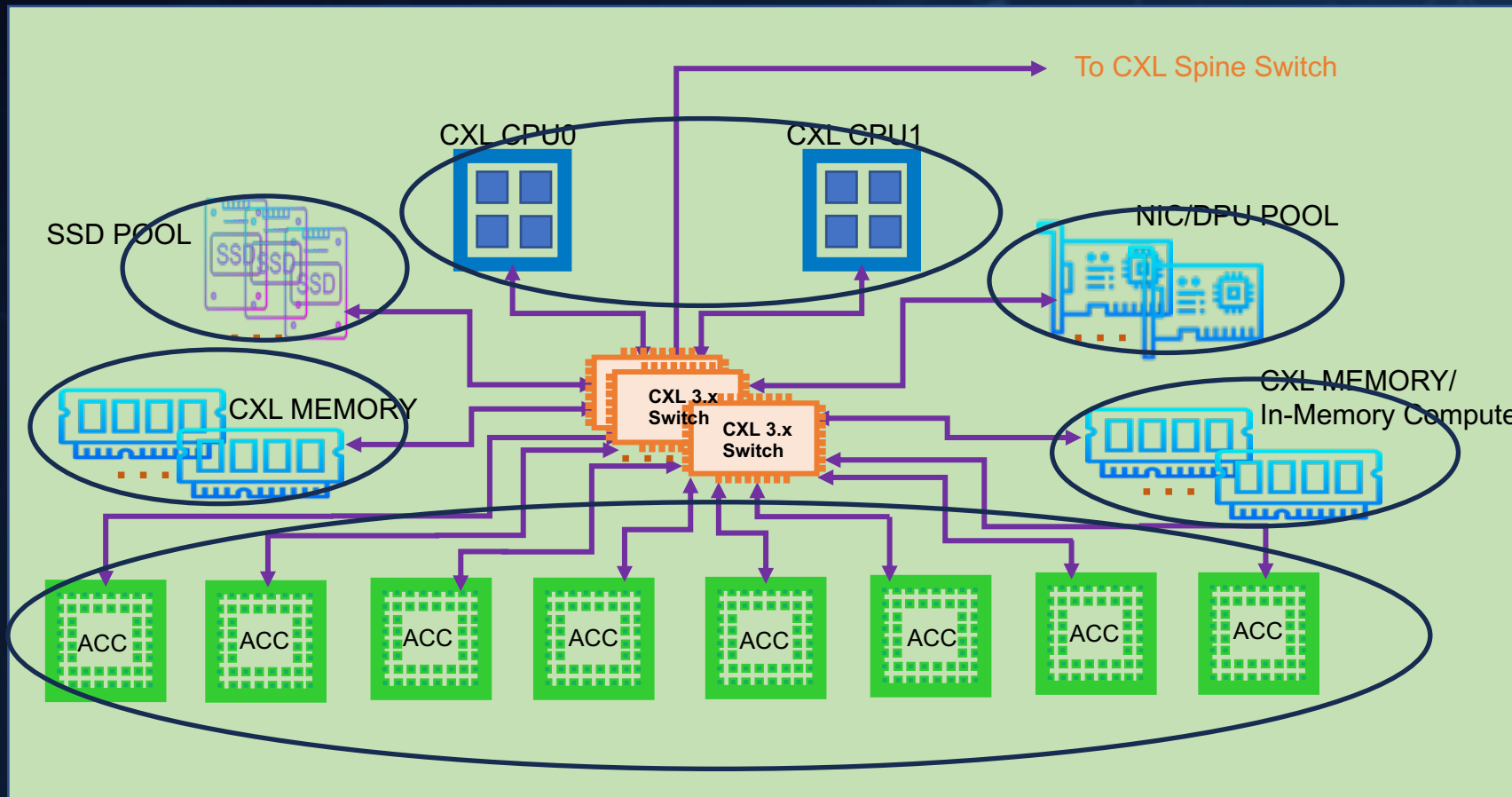


CXL 3.0 Fully disaggregation and composability of resources (rack-to-rack)



Source: CXL Consortium
CXL: Compute Express Link

CXL3.x/PCIe6.x Switch Enable Composable AI/ML Systems



- Memory pooling/sharing/expansion
- Supports All-to-All with scalable large switching capacity
- Fit for All-reduce, All-gather with super low latency and high bandwidth switching
- Scalable fabric network with up to 4,096 CXL devices
- Hybrid CXL/PCIe mode to connect CXL and PCIe devices
- Works with emerging CXL devices, e.g. In Memory Compute
- Lower total power consumption to reduce energy cost

Address:

1245 S. Winchester Blvd
San Jose, CA 95128

Web:

<https://www.xconn-tech.com>

Email:

JP.Jiang@xconn-tech.com



Thank you